JAIST Repository
https://dspace.jaist.ac.jp/

Title	Web上のHTML文書を用いた意外性のある情報の獲得支援
Author(s)	野口, 大輔
Citation	
Issue Date	2009-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/8100
Rights	
Description	Supervisor:東条敏教授, 情報科学研究科, 修士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

修 士 論 文

Web上のHTML文書を用いた意外性のある情報の獲得支援

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

野口 大輔

2009年3月

修 士 論 文

Web上のHTML文書を用いた意外性のある情報の獲得支援

指導教官　東条敏 教授

審査委員主査　東条敏 教授
審査委員　島津明 教授
審査委員　白井清昭 准教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

710056野口 大輔

提出年月: 2009年2月

Copyright © 2009 by Daisuke Noguchi

概 要

本稿では，Web 上の HTML 文書を解析した結果を元にした，意外性のある情報をユーザが得るための検索キーワードの想起を支援する方法を調査する．また，その具体的な方法として,「鳥式」ユーザインターフェースの改善と，トピックと関連語間の情報を元にした意外性評定の異なる二つの方法で，ユーザの「鳥式」で用いられる関連語における「意外性のある情報」の獲得支援を提案する．

実験では,「鳥式」ユーザインターフェースの改善について，システムの妥当性に対して 70%の肯定的な意見が得られた．トピックと関連語間の情報を元にした意外性評定について，提案した頻度や相互情報量，トラブル内順位を用いた手法で，Average Precision，R Precision において 23%の精度が最高 34%程度まで向上することを示した．

目 次

第1章　はじめに

1.1　研究の背景と目的

近年の情報化社会における急激な技術発展に伴い，我々一般のユーザがインターネットに触れる機会も日常的なものとなった．その中で，検索エンジンを用いて情報を得るということも，また有り触れた光景である．自らの欲する価値ある情報を，インターネット上に溢れ返る膨大な情報の中から迅速に取捨選択するノウハウは，ある種のリテラシとして認知されている程である．ここで，一般的に検索を行うという作業そのものは，ユーザ個人の知識に依存するところが大きいと言える．検索を行う際に入力するキーワードはもちろん，検索結果を絞るために用いるキーワードに関する語は，あらかじめユーザ個人の持ちうる知識をベースとした範囲に限定される．ユーザ個人の知るあるキーワードについて，例えばそのキーワードに関する問題回避や，あるいは行動に関する未知のアイディア，Tips について情報を求めようとする場合，ユーザが「意外」と思うような，ユーザの知識の範囲外であるキーワードを入力しなければならない場合がある．例えば,「頭が良くなる」等のキャッチフレーズでメディアに取り上げられた健康補助成分「DHA」が挙げられる．これは，過剰に摂取すると，血液の凝固作用を阻害し出血しやすい，あるいは出血が止まり辛い等といったトラブルを引き起こす要因となるものである．単純に「DHA」というキーワードで検索を行っても，有用な効能・効果を前面に押し出し紹介する Web サイトばかり上位に目立ち，出血のようなネガティブな情報を提供する Web サイトはほとんど見受けられない．健康のために普段から DHA のサプリメント等を購入している人でも，このようなネガティブな情報を知らないままに摂取し続けている状況は想像に難くない．しかし，あらかじめ検索を行う際に「DHA」に加えて「出血」という意外なキーワードを入力すると，検索結果上位に問題の事実が見つかる．ここで重要となる点は，このような意外なキーワードというものは，その関連がユーザの「知識の範囲外」，単純に言い換えれば「未知」であるがゆえに，何らかのシステムによって「気付かせる」ことが必要ということである．

このようなキーワードの想起を支援するため，我々の研究室では「鳥式」（図 1.1）と呼ばれる検索ディレクトリの開発を行ってきた．これは，ユーザが最初に入力したキーワード，つまりトピックに対して，関連語を意外な物まで含めて提示し，検索に利用できるようにする．なお，鳥式の第一の特徴は，鳥式が Web 文書に自然言語処理技術を適用することで自動生成されており，180 万語という大量のトピックをカバーしていることである．第二の特徴は価値ある情報を効率良く検索できるようにするため，いくつかの意味的カテ

図 1.1: 検索ディレクトリ「鳥式」

ゴリに属する関連語のみを提示することである．DHA の「出血」は「トラブル」というカテゴリ中の関連語として提示される．

現状で鳥式を使用するにあたり，いくつか既知の問題点がある.「鳥式」を用いて，あるトピックについて検索を行った場合，検索結果として得られる関連語はトラブル表現だけで数十に上る場合が多々ある.「鳥式」において検索された関連語は，表示順が Web 文書上でのトピックと関連語との共起頻度を元にしたスコアでソートされているのみであり，ユーザは全ての関連語に目を通さないと,「意外性のある情報」に辿り着けない可能性がある．また，全てに目を通したからと言って「意外な情報」に辿り着けるとは限らない．さらに，上述の上位概念の検索結果を継承する方法では．上位概念はより一般化されたトピックであるがゆえに，関連語の数が数十から数百へと飛躍的に増加する傾向にある．このような例では，全ての関連語に目を通すユーザは皆無であろう．また，関連語間の関係の傾向も不明瞭である．関連語相互の関係性も提示することにより，有用な情報を探し出すことがより容易になるであろう．

そこで，本研究ではユーザインターフェースの改善と，トピックと関連語間の情報を元にした意外性評定の異なる二つの方法で，ユーザの「鳥式」で用いられる関連語における「意外性のある情報」の獲得支援を提案する．ユーザインターフェースの改善については，平面グラフを用いてカテゴリ別に関連語を配置する．この時，トピックと関連語との共起頻度のみならず，単語間の類似度を用いて意味的に類似した関連語をまとめて表示する．トピックと関連語間の情報を元にした意外性評定では，トピックと関連語の Web 上の HTML 文書集合から計算した相互情報量や，トピックと類似度の高い語群の持つ関連語とトピック自身の持つ関連語を比較したデータを用い，ランク付けを行う．これらを統合的に用いることにより，ユーザの「意外性のある情報」の容易な獲得を促すことを目指す．

1.2 本論文の構成

本論文の構成は以下の通りである．第２章で分野の似た研究のみならず，本論文で用いた基礎研究について，関連研究として取り上げる．第３章でインターフェース部分と意外性評定方法に分けて提案手法について述べる．第４章ではそれぞれの提案手法についての評価方法とその評価結果について述べる．第５章ではまとめと今後の課題について述べる．

第2章　関連研究

本章では,「鳥式」インターフェースの改良に関してテキストマイニング可視化と類似語リストを，意外性評定方法に関して意外性の研究と，本論文で用いた既存の研究について取り上げる．

2.1　テキストマイニング可視化

SNS サービスやブログ等の近年の流行により，これまでは情報に対して受け身であった一般ユーザが，一転して情報を発信する側に回った影響も大きい．このような大規模かつ未加工な情報ソースの解析や調査を行う際に，それら全体の内容を把握するためには莫大なコストを必要とするという問題がある．この点について，高橋ら [1] は，テキスト中に書かれた評判情報や消費行動の抽出と可視化，マーケティングの分野への適用事例と今後の可能性について論じている．HTML 文書から書き手が書いた記事のみを抽出し，形態素解析，固有表現抽出・名詞句同定，評価表現抽出といった自然言語解析を行い，抽出した評価表現がどの対象物を評価しているかを見分けるために評価対を抽出し，評価対の出現頻度によるマイニングで精度を高める．これらによってテキスト情報から構造化された情報を抽出する．このように構造化された形にできれば,

- ある製品が世の中（ブログ全体の中）で得ている良い評価と悪い評価の割合
- 多くの人から言われている評価（e.g.「使いやすい」など）
- 他商品と比較したときの，評判の量や質（ポジティブな評価の割合など）

といった形で，ブログに含まれている評判情報を容易に整理し可視化することが可能であるとしている．可視化の例 (図 2.1) を見る限り，各商業施設に対しての客の印象が一目瞭然となっていることがわかる．

可視化によって，大規模なテキストデータから分析された情報（ここでは評判情報）をユーザに提示する点は本研究の趣旨と同じである．本研究では，さらに情報（ここでは関連語）間の類似度を用いて意味的に類似した情報をまとめて表示する手法を検討する．

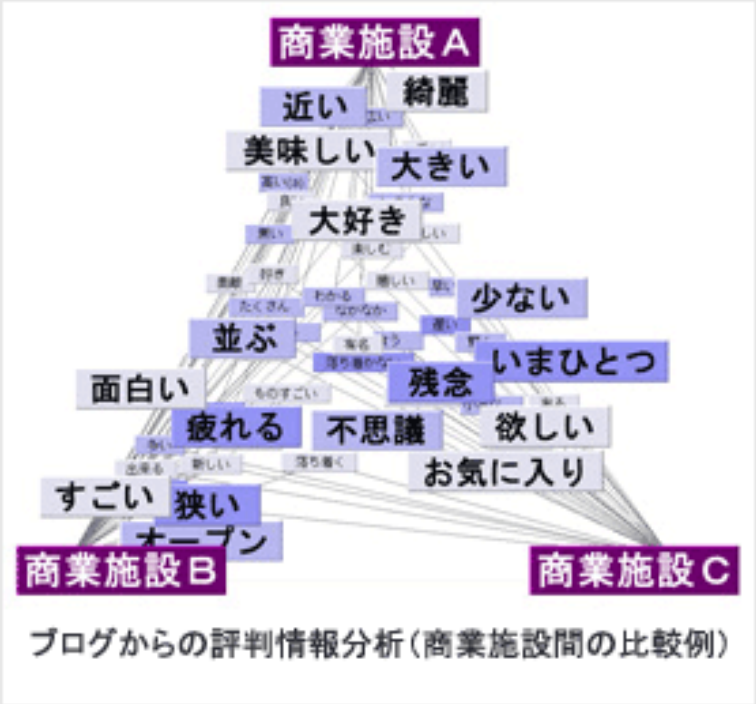

図 2.1: テキストマイニング可視化の例

2.2 類義語リスト

本研究では，インターフェースの改良に関して類義語リストを用いている．ここで用いた類義語リストは，係り受けの大規模な確率的クラスタリングの結果から，高精度で大規模な名詞の類似語リストを生成する方法について，風間ら [2] によって提案された手法によるものである．この研究では，クラスタリングの結果得られるクラス所属確率間の Jensen-Shannon ダイバージェンスを利用して語間の類似度を定義し，類似語リストを生成している．現時点で，語彙数を 100 万としたクラスタリングの実行に成功し，そこから 100 万語の各々の語に対して 500 個の類似語を類似度付きで生成することに成功している．

2.3 意外性のある知識発見のための Wikipedia カテゴリ間の関係分析

Wikipedia は，誰でも編集が可能な巨大なウェブ百科事典である．英語版 Wikipedia は 2008 年 8 月 11 日に 250 万記事，日本語版 Wikipedia は 2008 年 6 月 25 日に 50 万項目を超え，膨大な情報量を誇る百科事典として広く認知されている．

日本語版 Wikipedia は 9 個の主要カテゴリの下に，サブカテゴリ，記事が関連付けられており，大規模なグラフ構造を成している．各項目はそれぞれ複数の親カテゴリを持って

おり，また，同義語はリダイレクトとして関係付けられている．

Wikipedia の記事は，カテゴリシステムによってさまざまな観点からの分類がなされている．この特徴をうまく用いると，個別の記事からだけでは得られない意外な知識の発見につなげることができる．例えば,「麻生太郎」は「日本の内閣総理大臣」というカテゴリに属しているが，一方で「オリンピック射撃競技日本代表選手」というカテゴリにも属している．野田ら [3] は，このような意外な知識を Wikipedia から大量に発掘することを目的に，Wikipedia カテゴリネットワークに関する統計処理を行い，その結果を分析した．意外性の定義は人それぞれ異なるものの，評価結果とカテゴリ間関係の意外性にはある程度の相関がみられたとしている．

本研究では，Wikipedia を用いる方法ではなく，トピックと関連語という独自の視点に立って研究を行っている．

2.4 上位下位関係

本研究では，インターフェースの改良に関して上位下位関係を用いている．上位下位関係に関しては，一般の文書を知識源として，様々な上位下位関係の獲得手法が提案されているが，今回は隅田ら [4] が行った Wikipedia から高精度で大量の上位下位関係を自動獲得する手法について述べる．上位下位関係は，情報検索や Web ディレクトリなど，情報爆発時代の膨大な Web 文書へのアクセスを容易にする様々な技術への応用が期待されている．これまで，一般の文書を知識源として，様々な上位下位関係の獲得手法が提案されていた．しかしながら，概念具対物関係を含む広範な上位下位関係を獲得しようとすると，これらの手法では大量の文書を大規模な計算機資源を用いて処理する必要があり，例えば，隅田らが以前提案した上位下位関係の獲得手法を用いた場合，0.7TB の Web 文書を処理しても僅か約 40 万件の上位下位関係しか取れないなど，100 万件以上の上位下位関係を獲得するのは容易ではない．様々な事物に関する常識的知識をより密に記述する Wikipedia を知識源として，超大規模な上位下位関係データベースを「手軽に」構築することを目指した．具体的には，隅田らがこれまで開発した，Wikipedia の階層構造から上位下位関係を獲得する既存手法を拡張し，Wikipedia の定義文やカテゴリタグから獲得された上位下位関係候補についても，機械学習を用いて適切な上位下位関係を選別することで，Wikipedia 全体から高精度で大量の上位下位関係を獲得している．実験では，約 1.8GB の日本語版 Wikipedia から，約 188 万件の上位下位関係を 89.8%以上の適合率で獲得することができたと結論づけている．

2.5 トラブル表現の自動獲得

本研究では，意外性評定方法に関して，主に「鳥式」の意味的カテゴリ「トラブル」を中心に据えて提案を行っている．De Saeger ら [5] は，ある対象物（トピック）を用いる

際に関連する，潜在的トラブルや障害を発見する方法を提示している．この対象物とトラブル表現の関係の例としては,（薬，副作用）や（遊園地，身長制限）などのペアが挙げられる．De Saeger らの提案するこれらのトラブル表現の獲得方法は，３ステップから成る．まず最初に，上位下位関係を利用した語彙統語パターンや，DNV(Dependencies to Negated Verbs)，DAV(Dependencies to Affirmative Verbs) を用いて，Web 文書から学習データを集める．次に，SVM (Support Vector Machine) を使った教師あり学習で，トラブル表現と非トラブル表現を分類する．最後に，対象物と，上記で得られたトラブル表現を関連づけてペアにする．検証実験，並びにデータ収集には，大規模な日本語 Web 文書集合を用いている．トラブル表現を獲得する実験では，3 人の評価者が全員トラブル表現と判断したものを正解にした場合，精度 85.5%で 10,000 個のトラブル表現の獲得に成功し，対象物とトラブル表現のペアを獲得する実験では，3 人の評価者が全員トラブル表現と判断したものを正解にした場合，精度 74%で 6,000 の対象物とトラブル表現のペアを獲得できたとしている．

第3章　提案手法

本章では，インターフェース部分と意外性評定方法に分けて，具体的な提案手法を述べる．

3.1　「鳥式」インターフェースの改良

本研究では，HTML を用いて検索結果を表示していた「鳥式」について，検索結果を FLASH を用いた表示方法に変更し，ユーザのアクセシビリティ向上を促す．

3.1.1　FLASH を用いたユーザインターフェースの二次元グラフ化

「鳥式」ユーザインターフェースを，FLASH をベースにしたものに変更する．概要を図 3.1 示す．

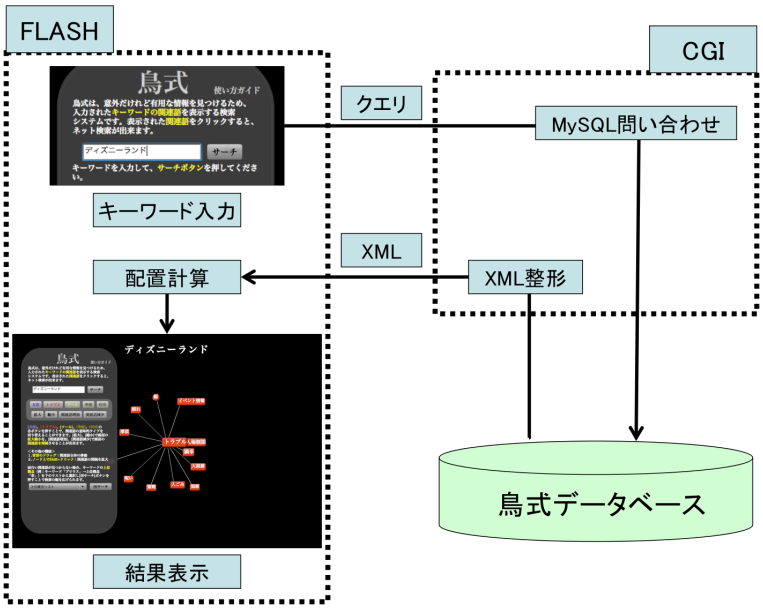

図 3.1: FLASH 版鳥式の動作概要

まず，ユーザが入力したキーワードを，鳥式データベース内に存在するかを問い合わせる．存在しない場合，システムはユーザに再検索を促し，存在する場合，該当するキーワードをトピックとしたデータを XML（Extensible Markup Language）形式のデータセットとしてシステムに返す．このデータセットの中には，後述の上位下位関係やノード配置に必要な頻度，類似度等といった情報が全て含まれている．これは，一連の計算を全てユーザのローカルマシンで行わせることを目的としたものである．これには，多数のユーザが同時にサーバにアクセスした場合に，計算負荷過多によるサーバダウンを防ぐ狙いがある．ただし，ローカルマシンの性能によっては，データの転送から表示までのプロセスに計算に因るタイムラグが生じる場合がある．本来の処理としては，全ての計算をサーバ側で行い，計算済みのデータセットをシステムに返す方法が望ましいが，このような設計を行った理由は，後述の実験で用いたサーバ群のスペックが貧弱だったため，大規模実験での負荷に耐えうるか曖昧だったためである．

得られたデータセットを解析し，後述のアルゴリズムに基づいて画面上に配置していく．配置には，主に極座標の概念を取り入れ，半径 r について共起頻度を，角度 θ について類似度を用いる．

トピック「ディズニーランド」について検索を行った例を図 3.2 で示す．

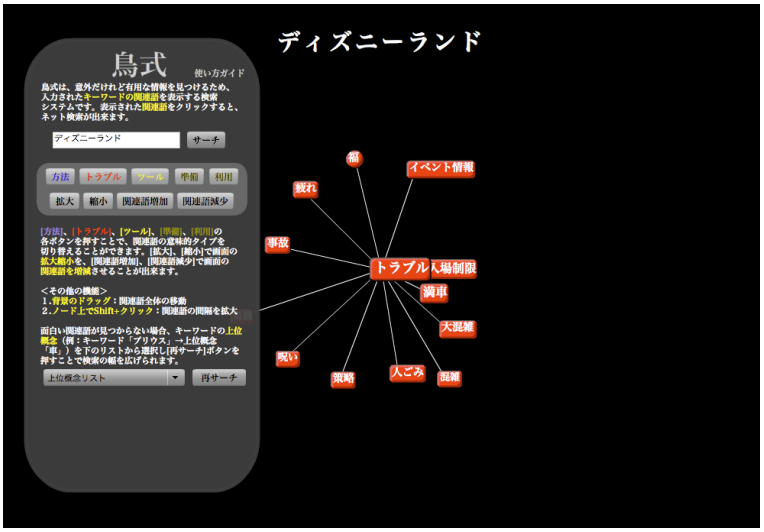

図 3.2: 検索結果表示

3.1.2 共起頻度に着目したノード配置

グラフのノード配置は，既存の意味的カテゴリである「トラブル」,「方法」,「ツール」等に分かれている．各々の意味的カテゴリに分けて，関連語を配置していく．ここでは，意味的カテゴリと関連語との距離 r は，トピックと関連語が Web 文書中に現れる際の共

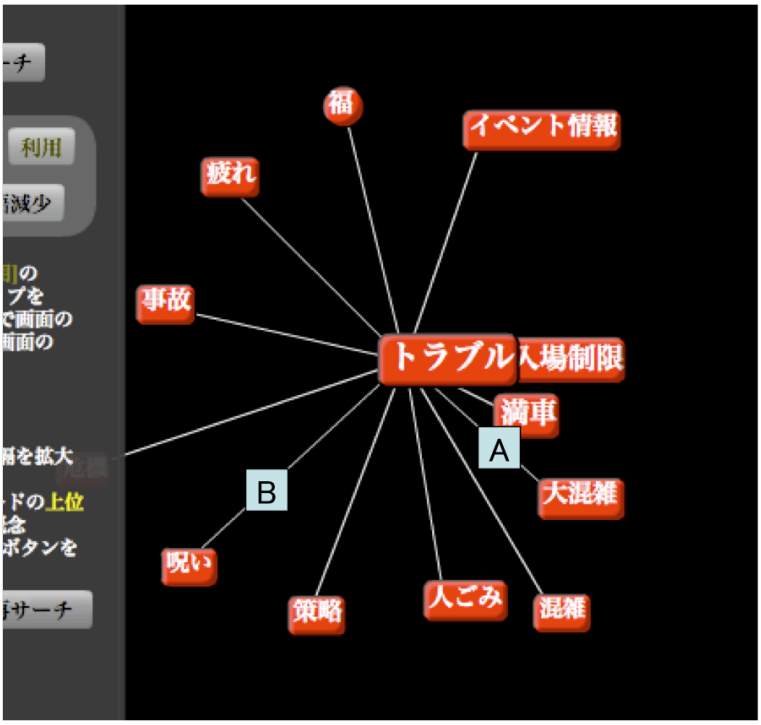

図 3.3: 中心-ノード間距離と共起頻度の関係

起頻度を元に正規化したスコアを用いている．ここでは，中心に近い程，その関連語がトピックとの共起頻度が高いことを表している．図 3.3 の A と B の例では，中心（トラブル）からの距離 r が $B > A$ となっているので，A の方がよりトピックとの共起頻度が高いことを表している．関連語の配置順，あるいは角度は，後述する．

検索直後において，初期配置されるノードは共起頻度の高いものから順に表示されるが，上限値が設定されている．これは，一度に大量にノードを表示する際に掛かるマシンへの負荷を軽減すると共に，共起頻度の低いものは有用でない情報が多いという経験則に基づいて実装された仕様である．

3.1.3 類似度に着目した擬似的ノードクラスタリング

中心からの距離は共起頻度を元にしているが，配置角度は類似度 [2] に基づいて決定している．ここでの類似度は，任意の単語と単語の間の類似度に比例してスコアを与えられたものである．配置角度の計算に先立って,「トラブル」,「方法」,「ツール」の意味的カテゴリについて，共起頻度の高い物から，各々最大 50 の関連語をサンプリングする．そして，以下の試行を各々の意味的カテゴリで行う．

1. 最もトピックとの共起頻度の高い関連語を角度 0 度に設定する．

2. 残った関連語の中で，1. と最も類似度の高い関連語を

$$\theta = \frac{2\ \pi}{n} \tag{3.1}$$

（n は参照した関連語の総数）に設定する．

3. 残った関連語の中で，2. と最も類似度の高い関連語を

$$\theta = \frac{2\ \pi}{n} * 2 \tag{3.2}$$

に設定する．

4. 以下，参照している関連語がなくなるまで同様の操作を繰り返す．

5. サンプリングされなかった関連語は，共起頻度の高い順に，既に配置済みである関連語群の中で類似度の高い上位２つの関連語の鋭角の角度の中心を取った角度に配置する．

このようなアルゴリズムを用いて，全ての関連語について θ を決定し，表示領域に描画を行う．以上で求めた角度 θ，並びに前小節で求めた半径 r は変数に格納しておき，次回以降の表示の際にはそれら参照することにより計算を省略する．

図 3.4 は，トピック「ダイエット」の対処に利用できるツール/材料を,「トマト」「砂糖」「風船」のような意外なものも含め提示した例であるが，意味的に類似した関連語がまとまって表示され，欲しい関連語を探すことを容易にしている．

3.1.4 上位下位関係を応用した検索拡張

意外性のある情報の獲得支援として，上位下位関係を用いた上位概念（例：トピック「ディズニーランド」に対する「遊園地」「テーマパーク」）による関連語の「継承」を実装している．(図 3.5）上位概念の継承とは，いわゆるオブジェクト指向言語におけるスーパークラスとサブクラスの間に発生する継承の概念と似ている．ここでの継承とは，上位概念（スーパークラス）の持つ関連語を全て下位概念（サブクラス）が受け継ぐことを指す．上位概念から継承を行うことで，追加された新規の関連語群をトピックが持つ既存の関連語群と合わせ，前述のノード配置アルゴリズムを用いることで，関連語間の関係性を視覚的に拡張し，ユーザが意外性のある情報を獲得しやすくするというイノベーション支援を行っている．

上位概念は，Web 文書から自動的に獲得されたものを大量に，トピックと対応する形でデータベースに保持しており，図 3.1 の CGI から渡される XML データセット内に上位下位関係のデータも含まれている．図 3.6 にて，トピック「ディズニーランド」に加えて，上位概念「遊園地」を継承させて検索を行った例を示す．図 3.2 と比較して,「ディズニーランド」固有のトラブル以外に「遊園地」に付随するトラブルも表示させ,「ディズニー

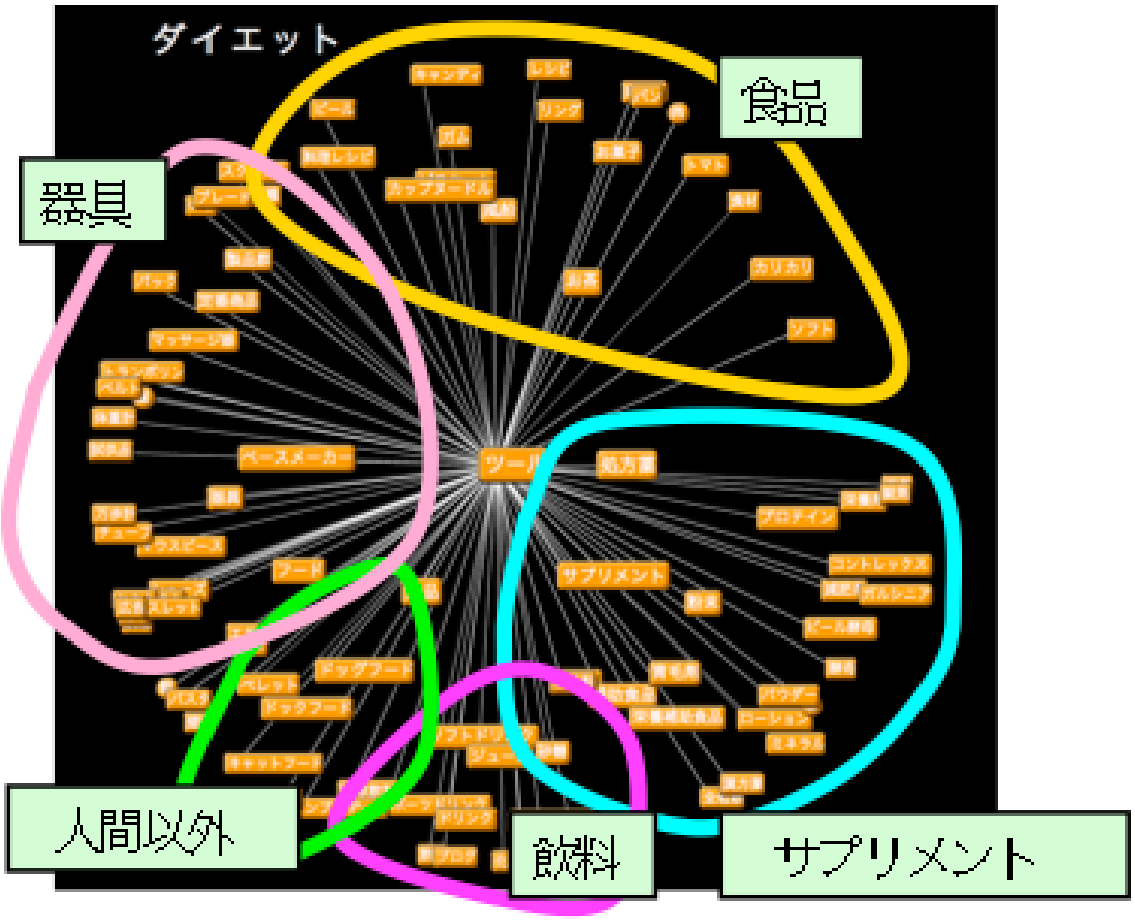

図 3.4: 「ダイエット」の「ツール」カテゴリのクラスタ例

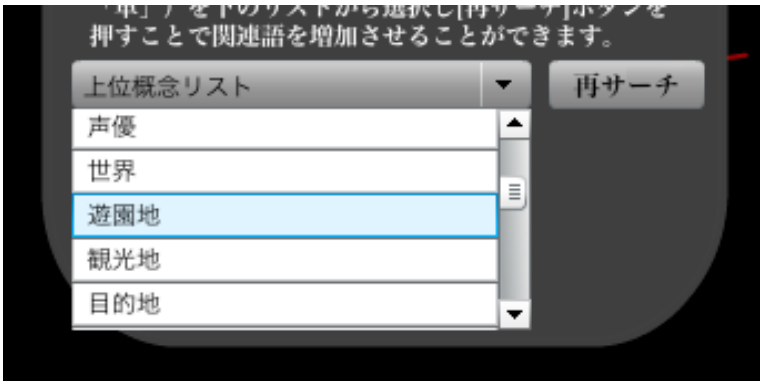

図 3.5: トピック「ディズニーランド」の上位概念リスト

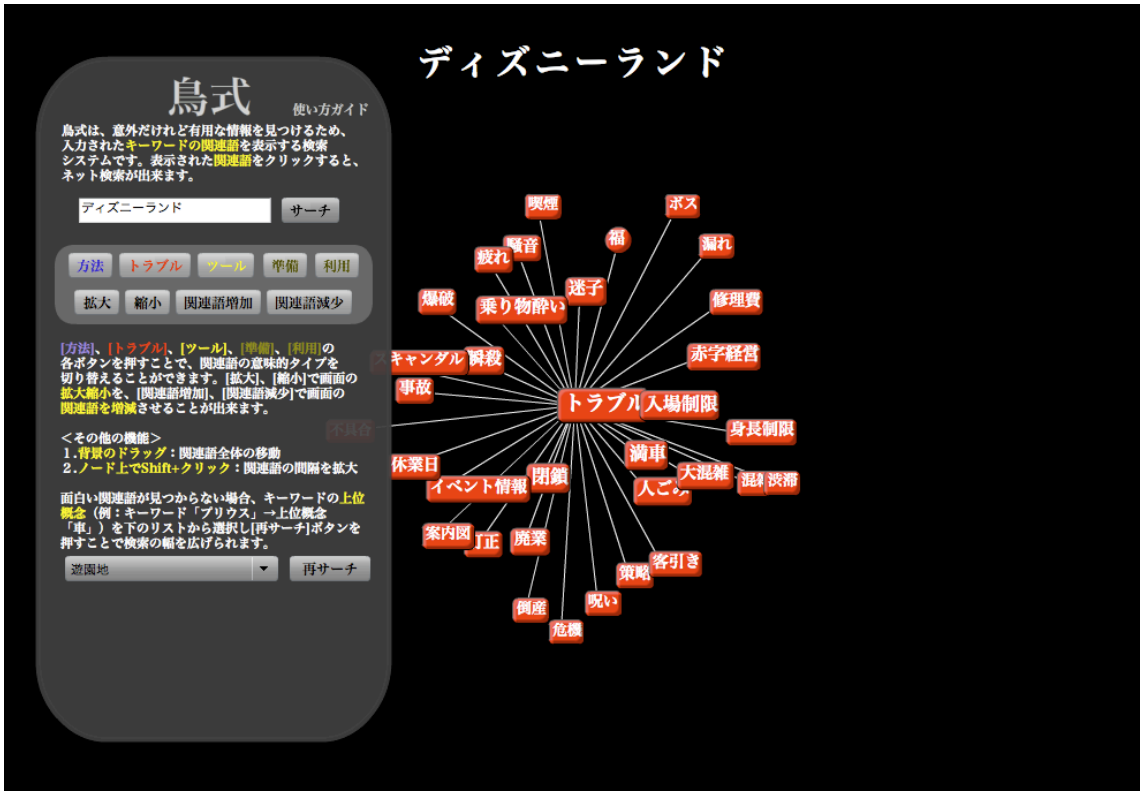

図 3.6: 上位概念を継承した例

ランド」のトラブルとして表示されていなかった「身長制限」などが新たに加わり，ユーザが意外性のある情報を獲得するための一助となっている．

さらに，昨年はじめに話題になった農薬ぎょうざ事件を例にとると，これまで開発した手法では，一昨年，つまり，事件以前の Web 文書から，トピック「ぎょうざ」の関連語として「農薬」を直接認識することはできなかった．しかしながら,「ぎょうざ」の上位概念になる可能性のあるものに「冷凍食品」があり,「冷凍食品」のトラブルとして「残留農薬」が認識できていることから,「残留農薬」を「ぎょうざ」のトラブルとして提示できる．つまり，騒ぎになる以前にぎょうざ事件をあたかも「予測」していたことになる．実際にぎょうざに付着していたものが「残留」農薬なのか意図的なものであるのかは今もって不明であるが，問題のぎょうざに関わった人々に「残留農薬」の可能性が事件の早い段階で示唆されていたとすれば，状況は改善されたかもしれない．鳥式はトピックに対して関連語を提示するという一見単純な処理しか行わないが，このぎょうざの例などは，そのような単純な処理ではあっても実社会でインパクトを持ち得ることを示唆しているものであると言えるのではないだろうか．

3.2 トピックと関連語間の情報を元にした意外性評定方法

3.2.1 「意外」とは何か

ここまで「意外」と一言で述べてきたが，一般に「意外」の意味するところは非常に曖昧であるので，ここで少し整理をしてみたい．人によっては例えば,「トラブル」という言葉を意識に上らせた瞬間,「可能な (タイプとして既知の) トラブル」(e.g., 肝炎) はすべて「意外でない」と言うかもしれない．しかしながら，我々は，こうした立ち場はとらない．一言で言えば，検索エンジンを前にして，ユーザが検索キーワードとして実際に入力はしないが，実は重要で価値がある関連語を，意外であると見なす．特に,「トラブル」のような一般的なキーワードを想起することは簡単であるが，あるトピック (e.g., アガリクス) に関するトラブルとして「肝炎」のような詳細度の高い具体的なキーワードを網羅的に想起し，検索キーワードとして入力することは容易ではない．ここで述べている意外の関連語としては，そうした詳細度の高い具体キーワードがまずは候補として挙げられる．これらをある程度網羅的に提示して，ユーザに価値ある情報にアクセスしてもらうのが鳥式の狙いである．

より具体的に，我々が意外であると考えている関連語としては，以下のようなものがある．

1. そもそも未知の関連語
2. 存在は知っていたが，トピックワードと重要な関係があるとは思っていなかった関連語

3. トピックワードと重要な関係があるとは薄々思っていたが，わざわざ検索をしてみようとは思い至らない関連語

ここで，以上のものの具体的な例をあげてみたいが，そもそもある関連語が意外であるか否かは，ユーザの知識，経験に依存する．したがって，以下では本論文の著者の視点で例をあげるしかないが，例えば，1 のタイプの例としては,「ソニー」の関連語でトラブルと分類されて提示される「パープルフリンジ（輝度差の高い被写体を撮影したときなどに、高輝度部分の周りに紫の縁取りができる現象)」がある．著者はソニー製の電化製品を購入した経験はあるが，ソニー製のデジタルカメラにパープルフリンジと呼ばれる問題がある（参照元はクチコミであるから，それが真実であるかは別として）ことは知らなかったし，パープルフリンジという言葉も直接は知らなかった．また，2 のタイプとしては,「ダイエット」のツールとしての「砂糖」がある．これは食間に砂糖を摂取することにより，脳に満足感を与えて無理なく食欲をコントロールするというものである．もちろん，著者は「砂糖」という語の意味するところは知っていたが，砂糖はダイエットをする上で障害となるものであり，それをツールとして利用するダイエット法があることは知らなかった．また，3 のタイプとしては,「skype」の「セキュリティー問題」がある．多くの skype ユーザは skype にセキュリティー問題があるかもしれないということは薄々思っているかもしれない．しかしながら，その内で実際にネット検索を行い，skype のセキュリティーについてチェックをしたことのあるユーザはまれであろう．ここでは，このような関連語も一種の意外であると見なす．仮に skype というキーワードを入力すると,「セキュリティー問題」が提示され，それをクリックするだけで有用な情報が得られるのであれば，それだけ skype のセキュリティー問題の具体的な情報に触れるユーザが増え，そこに有用性を見いだせるであろう．

ここで，1，2 のタイプの意外性は，万人に意外として受け入れられるであろうが，一方で 3 のタイプを「意外」と呼ぶことに抵抗がある人は多いものと思われる．しかしながら，今の文脈において重要なのは，検索という行為，そしてその瞬間におけるキーワードの想起である．つまり，検索という行為をしている瞬間に，検索キーワードとして想起されていないが実は価値あるキーワードというのが重要なのである．このような立ち場に立てば，3 のタイプもやはり意外と捉えるべきであり，このような意外性も提示することに意義があることになろう．また，このような意味での意外のキーワードは，実は広範囲に渡り，これらをカバーするためには，鳥式で行っているようにある程度網羅的に関連語を表示することに意義があるということになる．

さらに一点重要なポイントを挙げると，これまでは暗に一般ユーザの情報収集の支援という観点で説明を行って来た．しかしながら，類似したニーズは企業サイドにもあるはずである．例えば，企業にとって，ネット上で議論されている自社製品の問題点をチェックするのは，もはや必須であろう．当然，意外の問題点もチェックを行う必要があり，鳥式はそのような状況においても有用であると考えている．現在のシステムでも例えば,「XBOX」に対して「傷」といったものが提示されている．(ネット上で話題になっている XBOX の傷は，XBOX 本体の傷ではなくて，XBOX に挿入された DVD 等の傷である．) このよう

な関連語を特に外部から示唆されることなく，企業の担当者が思いつくのは，往々にして困難であり，やはり，鳥式のようなシステムである程度網羅的に列挙する必要があると考えている．

3.2.2 トピックと関連語の頻度情報に着目したランク付け手法

意外である，という判断はユーザの知識，経験に依存すると前節で述べた．一般的に，ここで言われているユーザの知識や経験は，Web 文書の情報の量と比例していると考えることができる．つまり，世間一般で知名度が高い情報ほど，それに関して記述がなされている文書数が多く，あまり知られていない情報（つまり，意外な情報）ほど文書数が少ないということである．文書数が多いということは，言い換えると Web 文書全体を参照した際にトピックや関連語が出現する頻度が高い，ということである．$f_{(o-t)}$ をトピックと関連語のペア $< o, t >$ が

$$o \text{ の } t \tag{3.3}$$

の形で Web 文書集合中に出現する共起頻度，$f_{(t)}$ を関連語 t が Web 文書集合中に出現する頻度とすると，

$$SCORE_{f_{(o-t)}} = f_{(o-t)} \tag{3.4}$$

$$SCORE_{f_{(t)}} = f_{(t)} \tag{3.5}$$

単純な手法として，$SCORE_{f_{(o-t)}}$，$SCORE_{f_{(t)}}$ がそれぞれ低い程，意外な情報が含まれる可能性が高いことが予測できる．

3.2.3 相互情報量に着目したランク付け手法

前小節の理論に従うとするならば，頻度の低い表現が上位に現われることになると予想されるが，共起表現においては頻度が低いからといって，必ずしも意外な表現とは限らないので，相互情報量を適用し表現を絞り込む．相互情報量により，トピックと関連語の結び付きの強さを評価する．相互情報量の値により順位づけを行うことによって，出現頻度に関係なく，結び付きの強い言い回しが得られるようになる．

$f_{(o-t)}$ をトピックと関連語のペア $< o, t >$ が o の t の形で Web 文書集合中に出現する共起頻度，$f_{(o)}$ をトピック o が Web 文書集合中に出現する頻度，$f_{(t)}$ を関連語 t が Web 文書集合中に出現する頻度，$f_{(o-t)sum}$ を全てのトピックと関連語の $f_{(o-t)}$ について足し合わせた値とすると，トピック o と関連語 t の間の相互情報量 mi は，

$$SCORE_{mi(o,t)} = \log 2(\frac{\frac{f_{(o-t)}}{f_{(o-t)sum}}}{\frac{f_{(o)}}{f_{(o-t)sum}} * \frac{f_{(t)}}{f_{(o-t)sum}}}) = \log 2(\frac{f_{(o-t)} * f_{(o-t)sum}}{f_{(o)} * f_{(t)}}) \tag{3.6}$$

と表す．相互情報量 $SCORE_{mi(o,t)}$ の高いものほど，そのトピックに固有の関連語であるということができる．トピック固有の表現は，希有な関連語とも言い換えることができ，未知の関連語である可能性が高いとも考えられる．

3.2.4 単語間類似度に着目した類義語によるランク付け手法

以下二つの小節で，前小節で得られた相互情報量にさらに手を加える．著者の主観であるが，トピック「あじさい」に対して，トラブル表現「毒」という関連語が意外であるとする．3.2.1 で述べた意外とは何かという話に通じる部分があるが，何故「あじさい」に対して「毒」が意外と感じるかという理由を考察すると，それは「あじさい」が観葉植物という観点から見て，観葉植物のような身近なものに人を死に至らしめるような「毒」が含まれていることが「意外」である，ということである．つまり，ここには観葉植物は安全という意識が前提となっている，と考えることが可能である．これはすなわち，あるトピック「X」が属するであろう集団「Y」には，トピック「X」のある関連語「T」は異端である＝滅多に存在しない，ということを表している．ここで，あるトピック「X」が属するであろう集団「Y」は，トピック「X」と類似度の高い語群と言い換えることが可能であると考える．

トピックと関連語のペア $< o, t >$ について，o と類似度の高いトピック n 個（類似度データベース [2] を参照する）の中で，t を含むトピックの総数を m 個とすると，

$$SCORE_{sim(o,t)} = SCORE_{mi(o,t)} * \frac{n}{m} \tag{3.7}$$

ここでは，$SCORE_{sim(o,t)}$ の高いものほど，類似度の高い類義語トピック群の中でも出現回数の少ない関連語が残る結果となり，より意外な情報である可能性が高いと考えられる．なお，o が類似度データベースに存在しない場合は評価データから削除する．

3.2.5 関連語内の順位に着目したランク付け手法

あるトピックと関連語のペア $< o, t >$ について，同一の関連語 t を持つトピックが 0 個以上存在する．そのトピック群の中で，相互情報量 $SCORE_{mi(o,t)}$ で降順ソートした場合に，上位に来るものは意外な情報が多いと考えられる．さらに，トピックと共起する数の少ない種類の珍しい関連語が上位に集まることになり，つまり,「そもそも未知の関連語」に属する組が上位に多くなると予想できる．

トピックと関連語のペア $< o, t >$ について，同一の関連語 t を持つトピック群の中で相互情報量 $SCORE_{mi(o,t)}$ で降順ソートした場合の $< o, t >$ の順位を k とすると，

$$SCORE_{rank(o,t)} = k \tag{3.8}$$

ここで，全てのトピックと関連語のペアで $SCORE_{rank(o,t)}$ を昇順ソートした場合，$SCORE_{rank(o,t)}$ の重複（$SCORE_{rank(o,t)}$ の値が同じ場合）が生じるが，その際は重複したトピックと関連語のペア群の中で相互情報量 $SCORE_{mi(o,t)}$ で降順ソートするものとする．

第4章　検証

4.1　「鳥式」インターフェースのGUI化

4.1.1　評価方法

提案手法を実装したFLASH版「鳥式」を構築した．

これを，本研究で開発を行っている「鳥式」も参画している特定領域研究「情報爆発時代に向けた新しいIT基盤技術の研究」プロジェクトにおける評価実験の一環として，研究者と学生で構成された評価者約60名に評価してもらった．

なお，評価者への質問は下記のとおりである．

質問1　表示された関連語の中に、あなたが通常の検索エンジンで検索しないけれど、有用なキーワードがありましたか？

質問2　関連語の意味的分類（トラブル、方法、ツール）は、あなたがよく検索するようなキーワードを含んでいますか？

4.1.2　評価結果

	はい	いいえ	はいと答えた割合
質問1	42	18	70%
質問2	35	26	57%

表 4.1: FLASH版「鳥式」評価結果

検索に成功した回数: 668回 (うち，27回は上位語を指定)
検索に失敗した回数: 663回 (うち，6回は人手で上位語を指定)

4.1.3　考察

検索回数をトータルで見ると，一人当たり平均20回で，予想以上にキーワードは投入されている．失敗した回数というのは，キーワードがデータベースに見つからなかった回

数である．検索を失敗した場合は，本来は上位語を入力することで上位語を継承した検索ができるのだが，その部分のインターフェースがうまくユーザを引きつけることができておらず，その機能はあまり利用されていないようであった．

質問に対する解答を参照すると,「表示された関連語の中に、あなたが通常の検索エンジンで検索しないけれど、有用なキーワードがありましたか？」とう問いに対して，70%の評価者が肯定的な解答を寄せている．評価者の研究者と学生は自然言語を学ぶ分野の方々であり，このような実験に対しては肯定的にバイアスが掛かっている可能性も否定できないが，上々の結果と言える.「関連語の意味的分類（トラブル、方法、ツール）は、あなたがよく検索するようなキーワードを含んでいますか？」という問いに対しては，57%の肯定的な意見と 43%の否定的な意見にほぼに二分されている．これは，普段の検索エンジン使用している状況を想定したものであるが,「鳥式」は「意外な情報」を引き出すために使われる状況を想定しているので，このような結果は妥当でもあるとも取れる．しかし，逆に言えば「意外な情報」の検索ばかりがニーズとしてあるとも限らない．理想としては，従来の検索エンジンでユーザが入力しているであろう関連語を表示しつつ，それに加える形で「意外な情報」を提案するシステムが理想だと言える．

また，システムに対するコメントとして，

- 説明が要らないくらい直感的にわかるデザインでないと，大衆的に広く使われにくい
- 検索速度の問題

等が挙げられていた．今回は付録 B のようなヘルプページを設けたが，実際に広く公開するとなると，デザインをシンプルかつ直感的なものにしなければならないだろう．検索速度の問題に関しては，データセットの容量を軽量化する，あるいは配置に置ける計算をシンプルにするなどといった工夫が必要である．

4.2 トピックと関連語間の情報を元にした意外性評定

4.2.1 評価方法

まず評価データを作成するため，新里ら [6] が収集した約 100,000,000 ページから抽出されたトピックと意味的カテゴリ「トラブル」の関連語のペア 2,379,155 組の内，ランダムに抽出したトピックとそれに付随するトラブル表現のペア 817 組について，作業者 12 名にラベル付けを行ってもらった．

今回関連語の中でもトラブル表現に限定したのは，作業者にとって「意外」と判断する基準として，意味的カテゴリ「トラブル」は比較的判断しやすい項目であるから，という点と，意味的カテゴリ「トラブル」「方法」「ツール」の中では,「トラブル」が経験的に最も意外性のある情報が見つかることが多かったためである．

ラベル付けは主に 3 種類あり，3 つのタスクから成る．ラベル付けに関して，作業者に配布したマニュアルの抜粋を付録 A に添付する．

要約すると，あるトピックとトラブル表現のペアについて，

1. 作業者にとって，トピックが既知であるか

2. 作業者にとって，その組み合わせは意外であるか

3. 作業者にとって，その組み合わせは他人は知らないかもしれないと思うか

の三項目を評価する．２.が主たる項目であるが，それ以外にも基準を設けた．１.はそもそもトピックについて知識が乏しければ，その関連語も未知であると判断する可能性が高いため，評価データとしては相応しくないと判断するために調査する．３.は知識レベルの高い人はどの組み合わせを見ても意外ではないと判断してしまうため，意外とする基準を引き下げる目的で設定したが，今回の検証実験では使用していない．

以上で作成された全817組のトピックとトラブル表現のペアについて，

条件１ 作業者６名以上が，当該ペアの組み合わせがWeb上に存在するとした

条件２ 作業者６名以上がそのトピックについて知っていた

の条件を適用し，評価データ全体とした．条件１はトピックとトラブル表現のペアがシステムにより自動的に獲得されたデータであるが故に，誤って取得されたもの（ゴミ）が多く見られるため，それらを除去する目的がある．条件２は前述の通り，そもそもトピックについて知識が乏しければ，その関連語も未知であると判断する可能性が高いため，評価データとしては相応しくないため除外する．

なお,「鳥式」で使用する状況を想定して，条件１.を除外したデータも評価用データとして別に用意した．

以上，二種類の評価用データについて，

- 作業者の少なくとも１名が当該ペアが「意外」だと判断した

を満たすものを正解データとして，提案手法計５手法にて，それぞれ評価を行った．正解データ数と評価データ数を，二種類の評価用データについて示したものが表4.2である．

	正解データ数	評価データ数
条件１・２	120	521
条件２	152	702

表 4.2: 正解データ数と評価データ数

上記で得られた正解データを元に，TOPnPrecision（トップからn個（nは上位半数の個数まで）データを取得した際の精度），Average Precision（トップから正解データを見つけたときごとの精度をもとめて、これらの精度を平均したもの），R Precision（トップからR個（Rは正解データの個数と等しい）での精度）の三種類を検討する．

4.2.2 評価結果

以下に，実験で得られた結果である TOPn Precision，Average Precision，R Precision を条件別に図 4.1 から図 4.6 に，Average Precision，R Precision の数値を条件別にまとめたものを表 4.3 に掲載する．表 4.3 中の (*) は，列中での最高スコアに対して付与している．

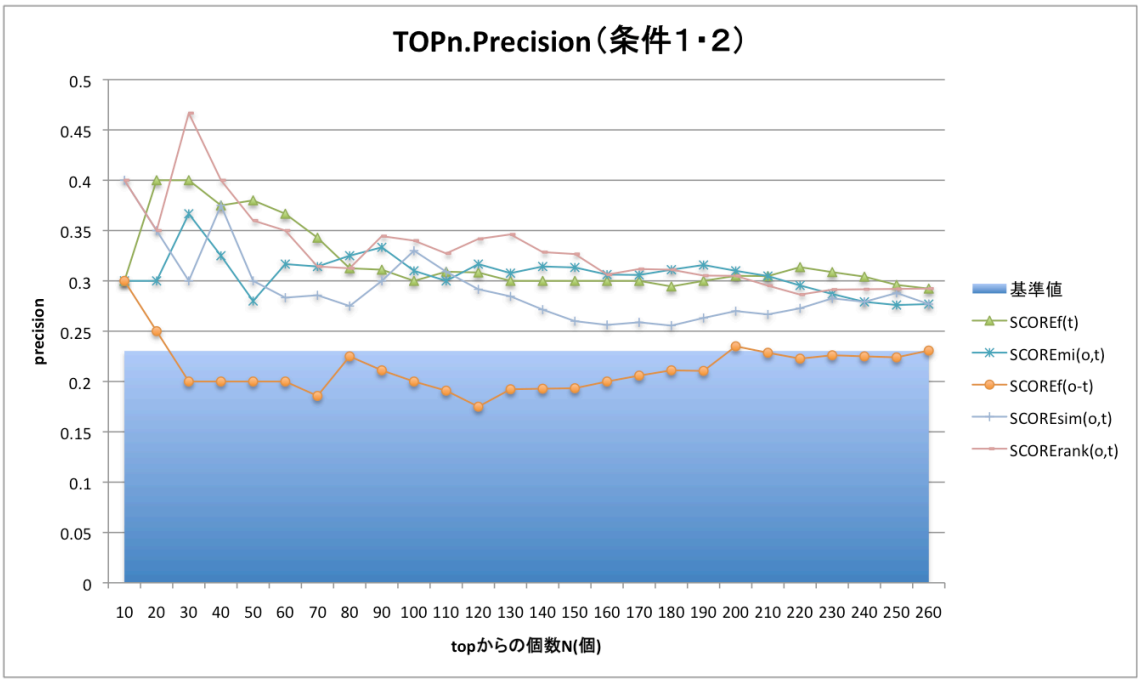

図 4.1: TOPn Precision（条件 1 ・ 2）

	条件 1 ・ 2			条件 2		
	基準値	Ave Prec	R Prec	基準値	Ave Prec	R Prec
SCOREf(t)	0.23033	0.30903	0.30833	0.21652	0.2856	0.28289
SCOREf(o-t)	0.23033	0.22975	0.225	0.21652	0.22269	0.21711
SCOREmi(o,t)	0.23033	0.29739	0.31667	0.21652	0.28738	0.28947(*)
SCOREsim(o,t)	0.23033	0.28447	0.29167	0.21652	0.26525	0.25658
SCORErank(o,t)	0.23033	0.33654(*)	0.34167(*)	0.21652	0.30023(*)	0.28947(*)

表 4.3: 条件別の Average Precision と R Precision

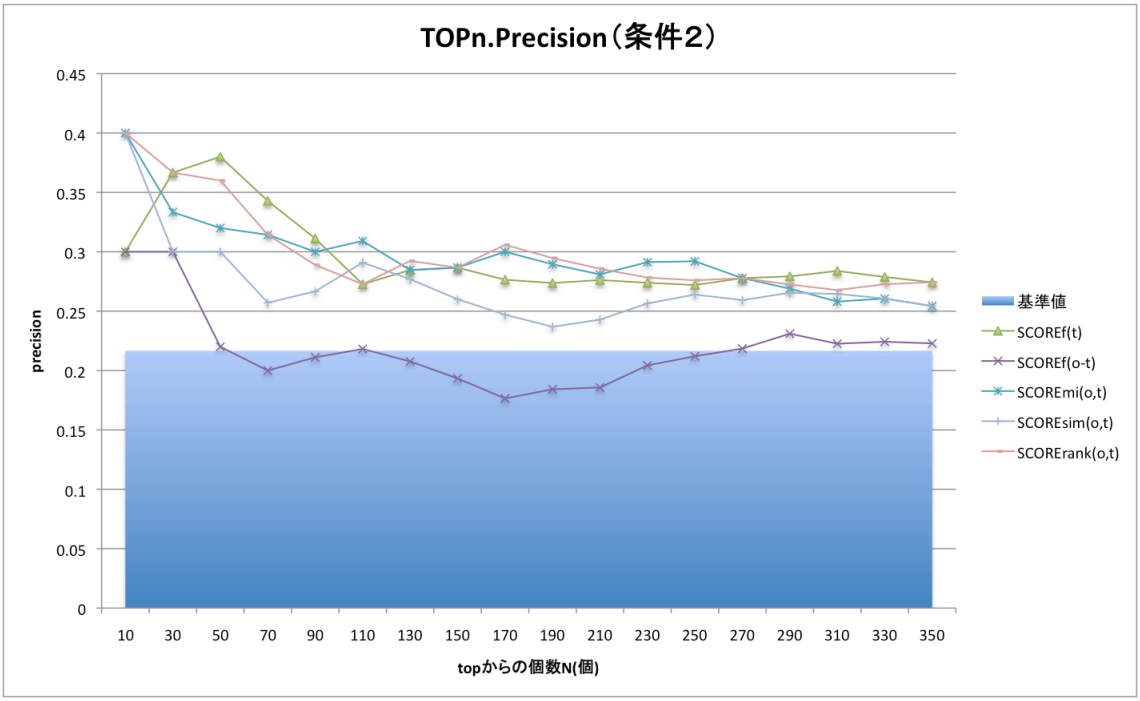

図 4.2: TOPn Precision（条件２）

4.2.3 考察

TOPnPrecision の結果（図 4.2，図 4.1）より，$SCOREf_{(o-t)}$ 以外全ての結果から上位に正解データが集まっている様子が確認できる．なお，何の手法も適用しなかった場合の精度である基準値も超えている様子が確認できる．$SCOREf_{(o-t)}$ で良い結果が出なかった要因としては，共起頻度の低いものが多数有り（e.g., 共起頻度２以下のものが 60%），ソートに寄る差別化がはかれなかったためと思われる．

さらに Average Precision，R Precision の結果から，$SCORErank(o,t)$（式 3.8）が最も良い結果となった．これは，トピックと共起する数の少ない種類の珍しい関連語が上位に集まることになり，つまり，「そもそも未知の関連語」に属する組が上位に多くなるとの予想が的中した形となったためと考えられる．各手法で得られた上位５個（ただし，被験者が意外としなかったものを除いた中での上位５個である）の意外な情報を表 4.4 に示す．それぞれの解説については，表 4.5 で行っている．

さらに，この表 4.4 を見ると，各々の手法によって得られた「意外な情報」とされたデータに違いがあることがわかる．上で述べた通り，$SCORErank(o,t)$ はトピックと共起する数の少ない種類の珍しい関連語が上位に来ており，トラブル表現の名称自体見慣れないものが多いのに対し，$SCOREsim(o,t)$ を見ると，トラブル表現自体は珍しくないにも関わらず，作業者が「意外な情報」と判断している例が多くあることがわかる．これは 3.2.1 でも述べた通り，

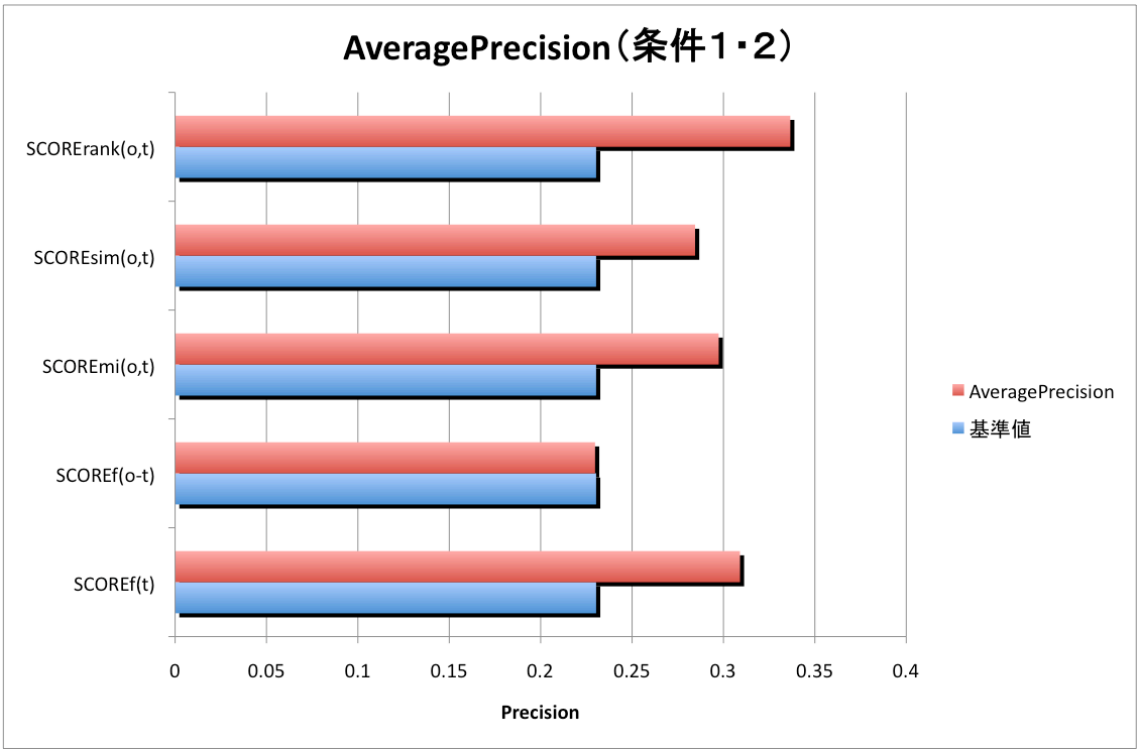

図 4.3: Average Precision（条件 1 ・ 2）

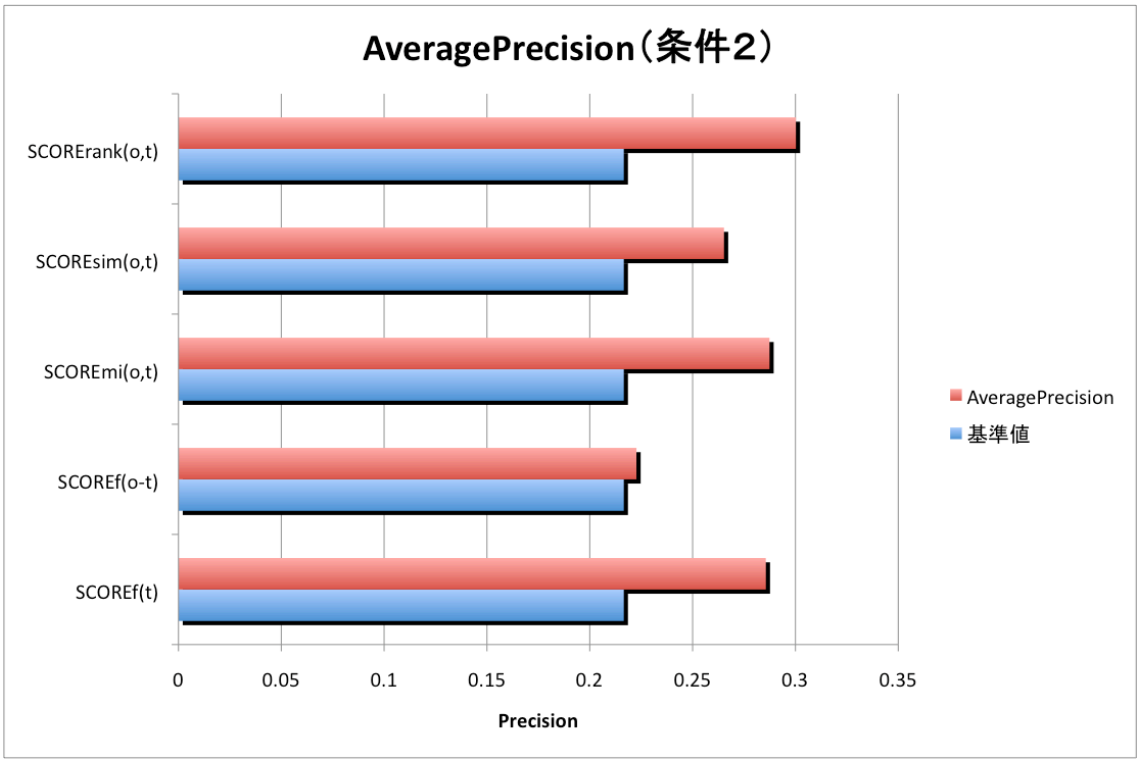

図 4.4: Average Precision（条件 2）

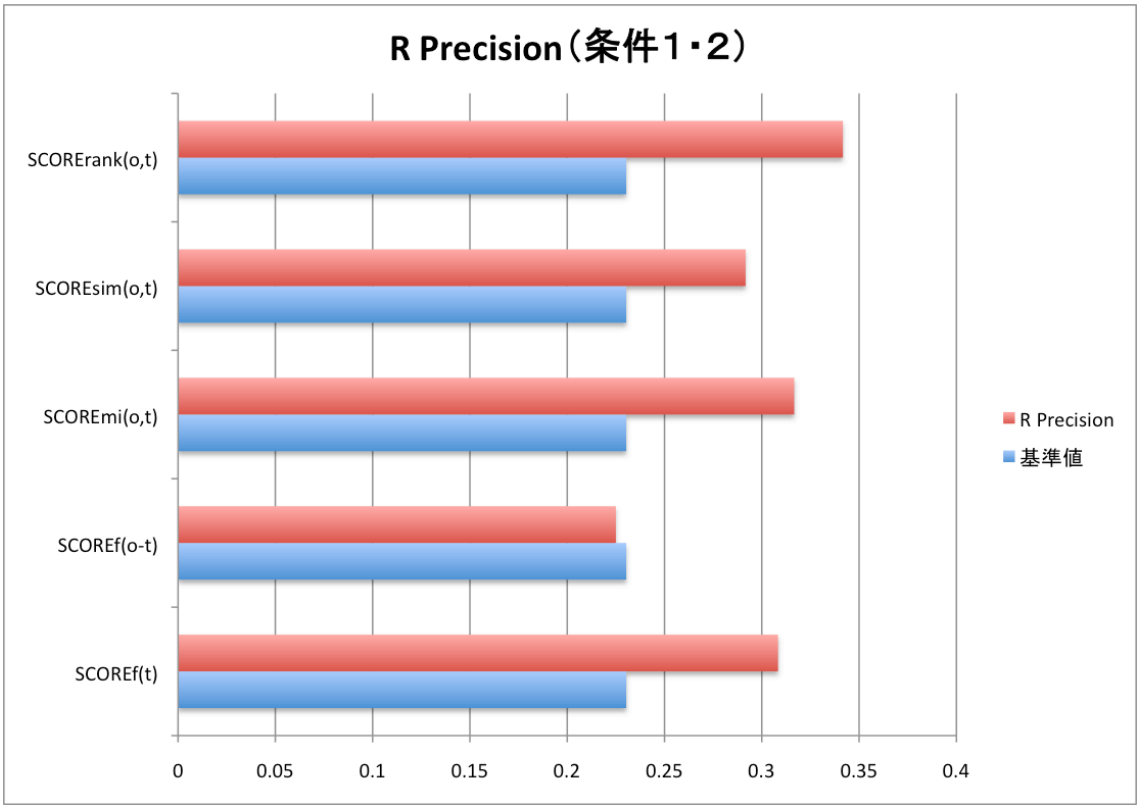

図 4.5: R Precision（条件 1 ・ 2）

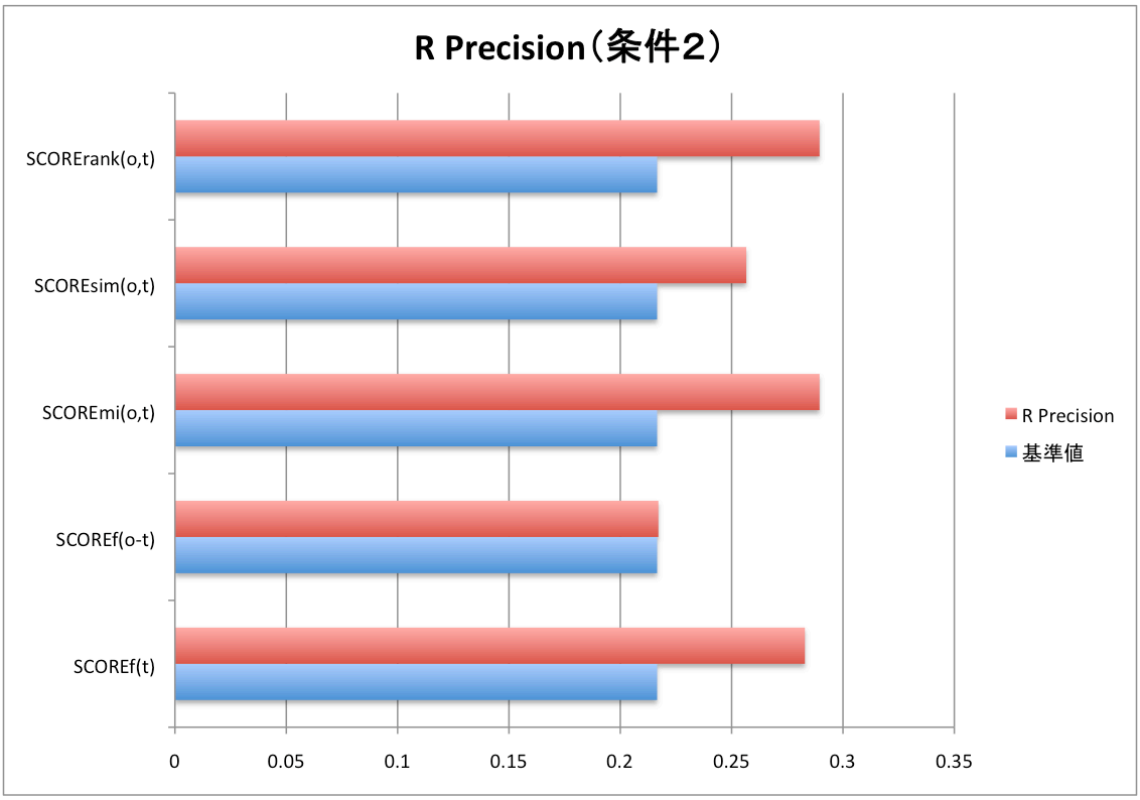

図 4.6: R Precision（条件２）

1. そもそも未知の関連語
2. 存在は知っていたが，トピックワードと重要な関係があるとは思っていなかった関連語
3. トピックワードと重要な関係があるとは薄々思っていたが，わざわざ検索をしてみようとは思い至らない関連語

の例の 1 と 2 でわかれているということである．これらのどれが「意外な情報」として最も価値があるかと問われれば 2 であると答えるべきである（かといって 1 に価値がないかと問われれば，NO と答えなければならないが）．今回の実験結果を見ると，$SCORErank(o,t)$ が最も精度が高く優れていると判断できるが,「意外な情報」の「質」で判断するとするならば，$SCOREsim(o,t)$ で得られた結果にも価値があると認めてもよいと考えることが可能である．

以上の結果より，精度自体はそれほど高くないものの，基準値を 10 %近く上回る結果を出したことから，少なくとも提案手法によって，何の手法も適用しなかった場合よりもいくらかの精度向上が実現できたと言ってよいだろう．

	トピック	トラブル表現
$SCOREf(t)$	ブータン	児童売買
	綿	農薬被害
	歴史劇	濫造
	商標権	希釈化
	蜂	糞公害
$SCOREf(o-t)$	歴史劇	濫造
	脳膿瘍	感染性疾患
	工場廃水	スラッジ
	陰陽師	暗殺
	ブータン	児童売買
$SCOREmi(o,t)$	歴史劇	濫造
	脳膿瘍	感染性疾患
	日本国歴代内閣	犯罪
	商標権	希釈化
	ブータン	児童売買
$SCOREsim(o,t)$	商標権	消滅
	カーボン	付着する
	樹林	乾燥化
	ブータン	民族問題
	カーボン	削りカス
$SCORErank(o,t)$	ひまわり	黒斑病
	ブータン	児童売買
	蜂	糞公害
	綿	チリダニ
	ひまわり	灰色かび病

表 4.4: 各手法で得られた上位 5 個の意外な情報

トピック	トラブル表現	解説
綿	農薬被害	綿への「枯葉剤」の付着
綿	チリダニ	綿ぼこり中のハウスダストの原因
歴史劇	濫造	中韓の高句麗論争
商標権	希釈化	「味の素」などへの商標の一般化
商標権	消滅	更新忘れによる権利の消滅
脳膿瘍	感染性疾患	細菌が血液によって脳に到達し，脳の中に膿が溜まる
日本国歴代内閣	犯罪	講和条約「日本国憲法」について
陰陽師	暗殺	暗殺者としての陰陽師
工場廃水	スラッジ	廃水処理殿物（専門用語）
カーボン	付着する	船舶に見られる「黒い汚れ」
カーボン	削りカス	車で，アルミリムの削りカスがカーボンリムを傷つける
樹林	乾燥化	サンショウウオ減少の要因
ひまわり	灰色かび病	専門用語
ひまわり	黒斑病	専門用語
ブータン	児童売買	インドの国境近くで児童売買
蜂	糞公害	養蜂場周辺での汚染

表 4.5: 獲得できた意外な情報の解説

第5章　おわりに

5.1　まとめ

本研究では,「鳥式」ユーザインターフェースの改善と，トピックと関連語間の情報を元にした意外性評定の異なる二つの方法で，ユーザの「鳥式」で用いられる関連語における「意外性のある情報」の獲得支援を提案した．ユーザインターフェースの改善については，平面グラフを用いてカテゴリ別に関連語を配置した．この時，トピックと関連語との共起頻度のみならず，単語間の類似度を用いて意味的に類似した関連語をまとめて表示した．トピックと関連語間の情報を元にした意外性評定では，トピックと関連語の Web 上の HTML 文書集合から計算した相互情報量や，トピックと類似度の高い語群の持つ関連語とトピック自身の持つ関連語を比較したデータを用い，ランク付けを行った．これらを統合的に用いることにより，ユーザの「意外性のある情報」の容易な獲得を促すことを目指した．このようなトピックと関連語に基づいた「情報の意外性」について研究した例は他にない．

実験では,「鳥式」ユーザインターフェースの改善について，システムの妥当性に対して 70%の肯定的な意見が得られた．トピックと関連語間の情報を元にした意外性評定について，提案した頻度や相互情報量，トラブル内順位を用いた手法で Average Precision, R Precision で最高 10%程度の精度が向上することを示した．

5.2　今後の課題

以下に今後の課題について述べる．

「鳥式」ユーザインターフェースの改善について，次のようなことが考えられる．

1. デザインをシンプルかつ直感的なものに
2. データセットの容量削減，配置に置ける計算の簡略化
3. ユーザへの提案の拡張

１，２に関しては，大規模な一般公開を行う際には必須の改善事項と思われる．ローカルベースでの配置計算も，サーバ側に負荷を分散可能なシステムが用意できるならば，サーバベースに切り替えることも視野に入れるべきである．３について，現在自動獲得した因

果関係を表示させたり（図 5.1），関連語の類似表現を用いたアナロジーを可能にするなどの機能が付加されている（図 5.2）.

図 5.1: 「うつ病」の原因（因果関係）を提示した例

トピックと関連語間の情報を元にした意外性評定についての今後の課題としては，次のことが考えられる.

1. 作業者の数を増やす
2. 評価用のデータを増やす
3. 「意外な情報」の基準をより明確する
4. 新たな手法の提案
5. 他の意味的カテゴリへの適用

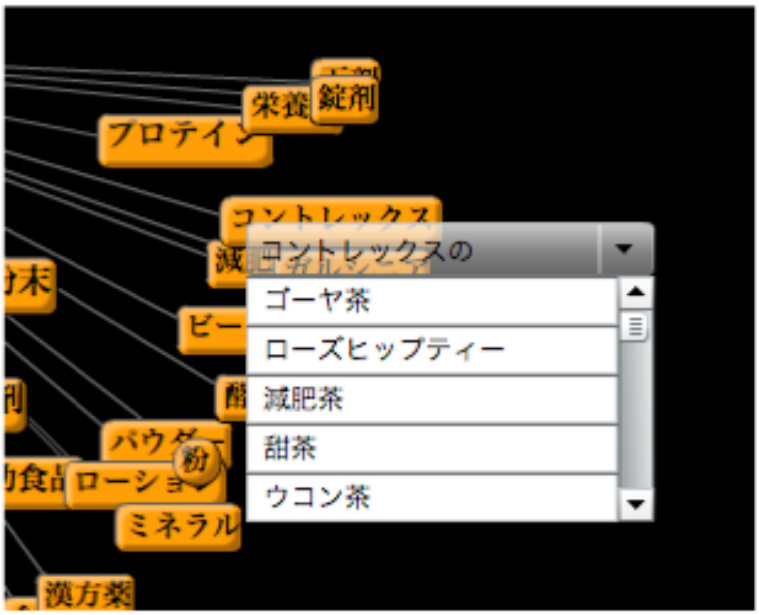

図 5.2: ダイエットのツールとしての「コントレックス」の類義語を提示した例

１，２については，データを増やし，各手法の信憑性をより向上させる．また，データ量が増えることにより，新たな「意外な情報」の基準が明確になる可能性もある．３について，作業者に一連の「意外性」を判定して貰う場合に，作業手順が,「意外性」という判断基準において様々な要因が絡み合い判定しづらいとの声が作業者から寄せられた．今回は何度も作業者と打ち合わせを繰り返し意識の統一を図ったが，これから作業者を新たに増やす場合など，判断基準となるマニュアルについて，もう少し吟味する必要性を感じた．４について，本研究で提案した手法を発展，あるいは相互に組み合わせて精度の向上を目指す．５について，本研究では「鳥式」の意味的カテゴリ「トラブル」に限定して研究を進めたが，今後は他の意味的カテゴリ「方法」「ツール」にも適用できるようにする．ただし，今回の手法はあくまで意味的カテゴリ「トラブル」に適応した場合，一定の精度向上が見られたものであり，同様の手法で「方法」「ツール」で同様の精度向上が見られるかは定かではないし，作業者による新たな大規模作業が必要となる．

A　意外性評定実験タスクマニュアル・抜粋

タスク 1.

ある名詞句のペア（A,B）に関して、A が既知であれば“ O ”、知らなければ“ X ”としてください。

Ex.

｜あじさい｜中毒 ||||　（｜は列の区切りと考えてください）

　　↓あじさいは知っている

｜あじさい｜中毒｜O |||

｜虻田町｜噴火 ||||

　　↓虻田町は……知らないな

｜虻田町｜噴火｜X |||

タスク 2-1.

次に、(A,B）に関して、A が既知である場合、“ まず検索をせずに ”「A のトラブル語として B は意外である」と思える場合に “ O ”、そうとは言えない場合は “ X ”というラベルを付けてください。

Ex.

｜あじさい｜中毒｜O |||

　　↓あじさいで中毒が起こるとは知らなかった。(この時点では真偽不明)

｜あじさい｜中毒｜O｜O ||

次に、上記で“ O ”のついたものだけ「“ トピック ” “ トラブル ”」として検索を行い、そもそも B が A のトラブルという関係が無かった、あるいは A と B が係り受け関係にない場合は“ I ”というラベルに変更してください。

Ex.

｜あじさい｜中毒｜O｜O ||

　　↓あれ？あじさいが中毒を治すという事例だった（例なので嘘です）

｜あじさい｜中毒｜O｜I ||

タスク 2-2.

A が未知の場合、A について検索を行い知識を得た上で 2-1 と同様にラベル付け行ってください。

Ex.

｜虻田町｜噴火｜X |||

　　↓有珠山が近くらしいし、おかしくない

｜虻田町｜噴火｜X｜X ||

タスク 3.

最後に、あなたが「A についてのレポートを書かなければならない」という状況を想定し、B というトラブルを見て調べてみたくなるかどうか、調べてみたいと思えば“O”、そう思わなければ“X”としてください。タスク 2 において既に“I”がついている場合は同じく“I”としてください。
Ex.
| あじさい | 中毒 | O | O ||
↓是非調べてみたい
| あじさい | 中毒 | O | O | O |

| 虻田町 | 噴火 | X | X ||
↓レポートを書くなら必要かもしれない
| 虻田町 | 噴火 | X | X | O |

B　検索ディレクトリ「鳥式」の使い方

鳥式は，意外だけれど価値ある情報を検索するためのシステムです．基本的には入力された検索キーワードに関係の深い単語を，トラブル，ツールといった意味的な固まりに分類した上で表示します．表示された単語をクリックすることで，その単語に関する文書が入手できます．(現在は Yahoo 経由で文書を入手するようになっています．)

鳥式は大量の Web 文書から計算機による自動的な学習で，自動的に生成されたもので，現在約１３０万語の検索キーワードをカバーしています．鳥式を用いて発見できる「ぎょうざ」と「残留農薬」というトラブルとの関係は，昨年度の Web データから自動で学習された結果で，ある意味，農薬ぎょうざ事件をシステムが予見していたことになります．すこし種明かしをしますと,「ぎょうざ」は「冷凍食品」の一種であり,「冷凍食品」の残留農薬の問題はだいぶ以前から話題になっていたため，この「予見」が可能になっています．つまり，システムに「ぎょうざ」をより一般的な「冷凍食品」という概念に抽象化させることで予見が可能になった，ということです．

なお，あくまで機械が自動的に学習した結果ですので，明らかに間違っている，あるいは意味をなさない単語も表示されますが，ご容赦ください．

では，鳥式の実際の使い方を具体例を踏まえながら順を追って説明していきます．

B.1　もっとも基本的な使い方

鳥式を開くと，図 3 のような画面が現れます．左側にあるグレーの部分をコントロールパネルと呼びます．基本的な操作は，このコントロールパネルにある検索窓に検索キーワード（全角 50 文字まで）をいれてから，その周りにあるボタンを押すことで表示方法を変えながら，関連語を探します．

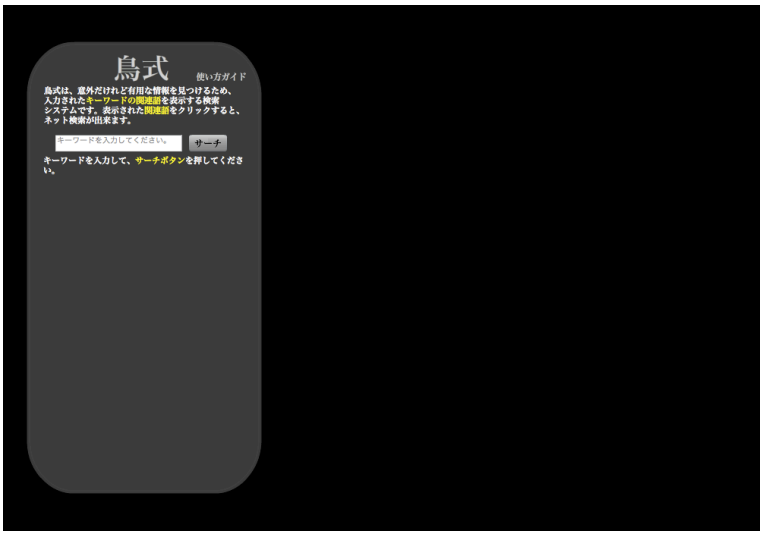

図 3: 鳥式基本画面

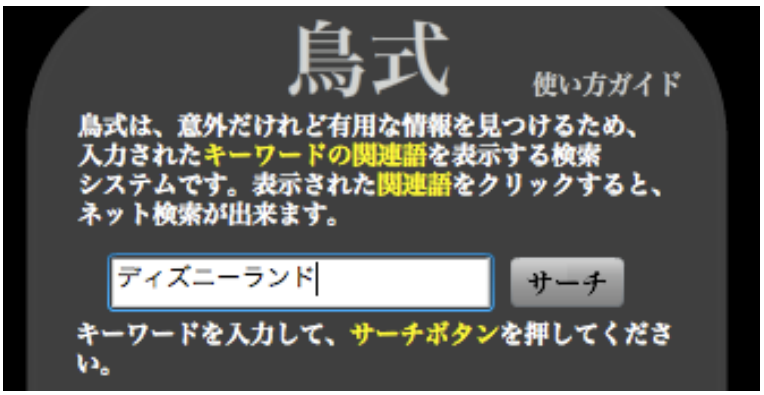

図 4: キーワード入力

では,「ディズニーランド」という検索キーワードを入れることを考えます.(図 4) 検索キーワードを入れた後，検索窓の右隣にある「サーチ」という名前のボタンを押します.

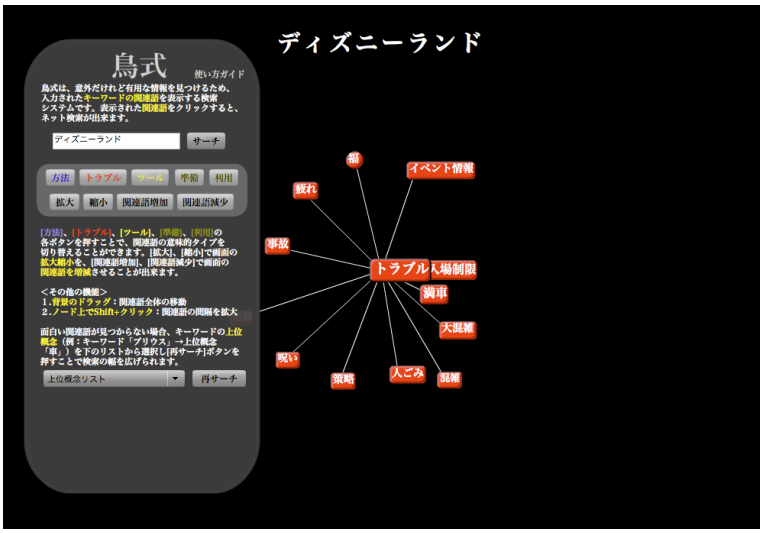

図 5: 検索結果表示

そうしますと，今度は図 5 のような画面が表示されます．ここで様々な単語が赤の四角の中に書かれたものの「群れ」が現れますが，これらが冒頭で述べた検索キーワード（今は「ディズニーランド」）に関連の深い単語です.

B.2　関連語の表示操作

現時点では各の単語は見やすいとは言えません．では，次にこれらのキーワードを見やすくするための操作を説明します.

まず，あなたがこれから「ディズニーランド」に出かけようとしていると想定してください．そのための準備としてうれしいことの一つは,「ありそうなトラブル」をあらかじめ Web でチェックしておくことです．鳥式では，まさにそういう状況での検索を楽にすることを狙っています.

ここで，ディズニーランドに行った際に出くわすと面倒な「満車」「入場制限」といったトラブルが見つかると思います.(図 6) ここで，例えば「満車」という単語をクリックしますと，Yahoo を経由して，ディズニーランドでの駐車場が満車であるという困った状況に関する文書が入手できます．おそらく，ディズニーランドに行くユーザで，事前に駐車場が満車であるかどうかを気にする人はそんなに多くないでしょう．このようなユーザにとって，鳥式は「転ばぬ先の杖」として機能します.

なお，この状態はすべての関連語が表示されている状態です．関連語の総数が多すぎて見辛いと感じた場合，コントロールパネルの「関連語減少」というボタンを押してくださ

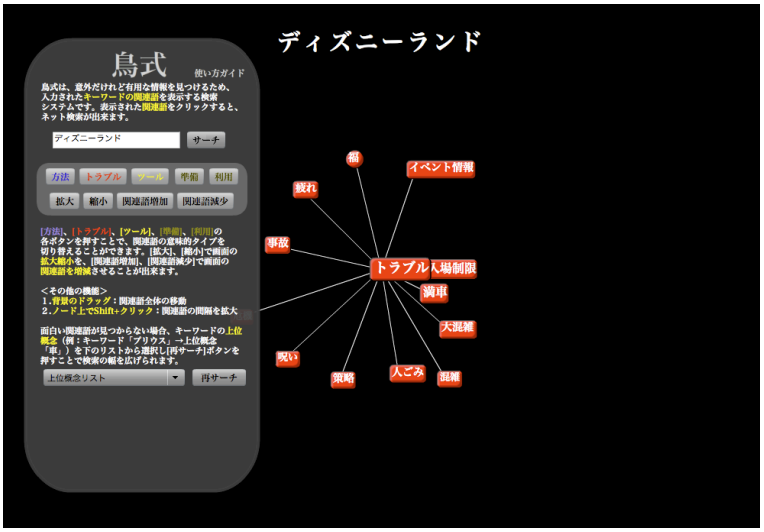

図 6: 関連語の操作

図 7: 関連語減少

い．そうしますと，表示される単語が減ります．(図 7) その他，もし，画面の外に単語が表示されるようでしたら，やはりコントロールパネルにある「縮小」というボタンを押してください．そうすることで，単語の群れが縮小表示され，画面外に表示されていた単語が見えるようになります．また，単語以外の部分を「ドラッグ」することで，視点を自由に移動させることも可能です．

B.3 「継承付き関連語」ボタンの利用法

これまでの操作では特別興味深い単語が見つからない場合があります．そのような場合は，入力された検索キーワードを，一段上位の概念に一般化し，それに関連する単語を表示させることができます．

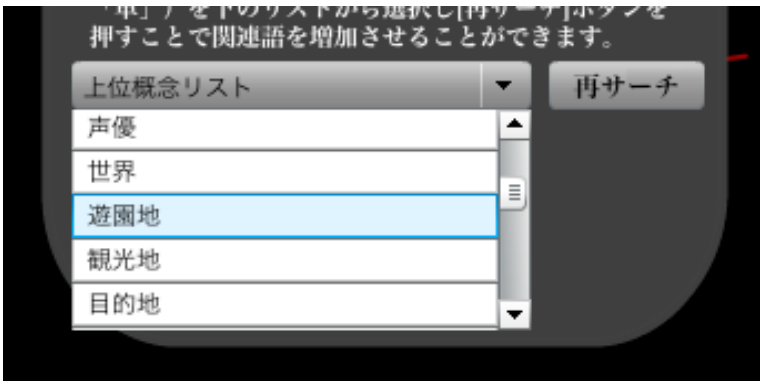

図 8: 上位概念一覧

とりあえず，検索窓に検索キーワードをいれて，関連語ボタンを押すと，コントロールパネルの上位概念リストに上位の概念の候補が並べられます．たとえば，ディズニーランドという検索キーワードを入れると，図 8 のような上位概念のリストが表示されます．ここで，例えば,「遊園地」というディズニーランドの上位概念をクリックし，上位概念のリストの上にある「再サーチ」というボタンを押すと，ディズニーランドだけではなく,「遊園地」というキーワードに関連の深いキーワードが表示されます．

遊園地のチェックを入れてから「再サーチ」を押すと，図 9 のようなものが表示されますが，関連の深い単語が増えていることがわかります．例えば，子供をアトラクションに乗せる時の制限となる「身長制限」や,「渋滞」「乗り物酔い」「迷子」といった単語がでてきます．これらをクリックすると，公式ページではなく,「ディズニーランド」の「身長制限のあるアトラクション一覧」「渋滞が回避できるルート一覧」「ディズニーランドで乗り物酔いするアトラクション」といった情報や,「ディズニーランドでは実は迷子のアナウンスはない」といった意外な情報まで入手できます．

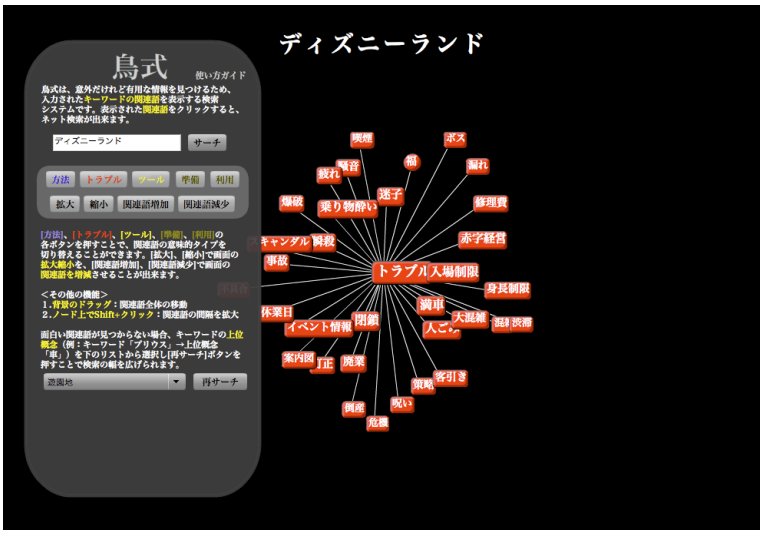

図 9: 継承付き関連語

B.4　鳥式がカバーしていない検索キーワードを検索する方法

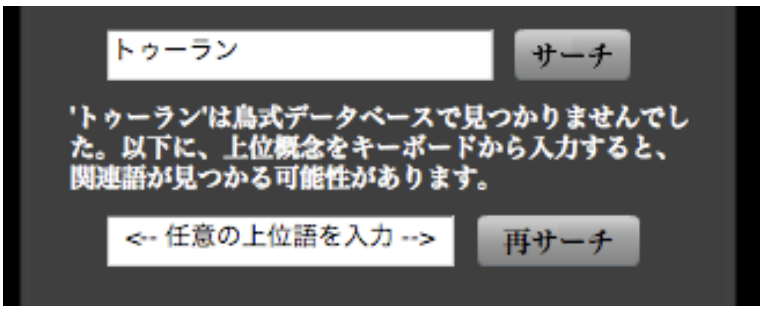

図 10: 上位語候補が見つからない例

現在鳥式は約１３０万語の検索キーワードをカバーしていますが，当然，カバーされていない検索キーワードが入力されることもあるでしょう．その場合は，上で説明した「再サーチ」ボタンをうまく使うことで，検索キーワードに関連の深い単語を表示させることができます．例えば，現在は,「トゥーラン」という車は，検索キーワードとしてはカバーされておらず，検索窓に「トゥーラン」と入力すると図 10 のようなメッセージがでます．

ここでトゥーランに関する意外な情報を入手したい場合は,「再サーチ」ボタンのすぐ左にある入力窓にトゥーランの上位概念である「車」を入力し,「継承付き関連語」ボタンを押します．そうすると表示されるのが図 11 ですが,「修理費」,「傷」といった事柄に関する情報を入手することができます．これらは，トゥーランの関連語としては得られていませんが、上位語の関連語をトゥーランの関連語としても利用でき、これらを表示することでトゥーランの意外な情報を入手する手がかりになりうると考えられます．

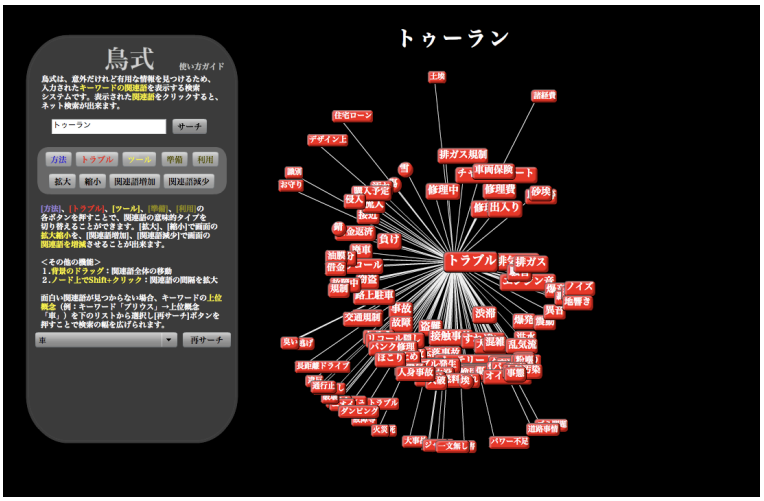

図 11: 任意の上位語を継承した例

B.5　トラブル以外の意味的カテゴリ

これまでは主として,「トラブル」という意味的カテゴリに属する情報を探す方法について述べてきました．ところが，鳥式のコントロールパネルを見ると分かるように，トラブル以外にも「方法」「ツール」「利用」「準備」というボタンがあります．以下ではこれらのボタンの意味，使い方について説明します．

鳥式を設計した際の基本的方針は，検索キーワードの「利用」に関係した様々な情報を見つけられるようにしたいということでした．これまで説明してきた「トラブル」も実際に検索キーワードを利用する際に問題となる「トラブル」を中心に提示するというのが基本的なアイディアです．

他のボタンについて説明しますと，まず,「方法」ボタンは，検索キーワードの利用にまつわる具体的な方法を提示するものです．例えば，図 12 は検索キーワードとして「鮎」を入力し,「方法」ボタンを押した後の表示ですが，鮎を利用する方法，つまり,「食べる」ための方法として,「塩焼き」「お茶漬け」「ムニエル」「釜揚げ」，意外なところでは「すき焼き」などの料理法，あるいは，利用する，つまり「食べる」前の準備段階での「釣る」具体的な方法としての「友釣り」，あまり知られていないところでは「ころがし」などの情報が入手できます．

ところで図 12 を見ますと，外周に単語群が集中して見辛い印象を受けます．このような場合，適当な単語を「Shift+クリック」しますと，その単語周辺の単語群が展開して見やすくなります．

また,「ツール」ボタンは，利用をする際に有用なツール，もしくは材料を提示します．図 13 は「ダイエット」という検索キーワードを入力した際の例ですが，ダイエットに用いるものとして「ビール酵母」や「プロテイン」など一般的に知られてる単語の他にも，「ガルシニア」といった見慣れない単語や，一見ダイエットの天敵となるような「砂糖」

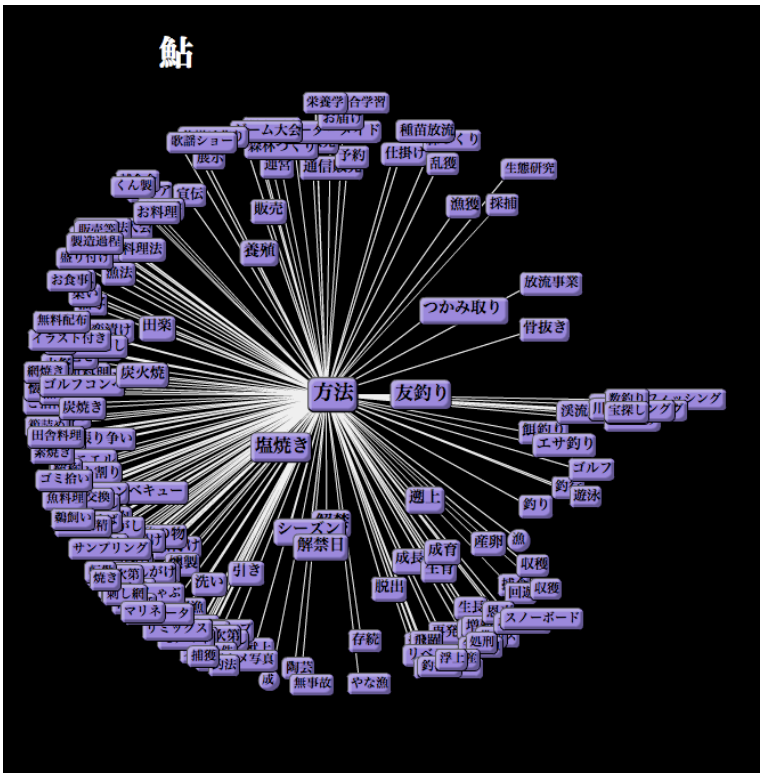

図 12: 鮎の「方法」カテゴリ

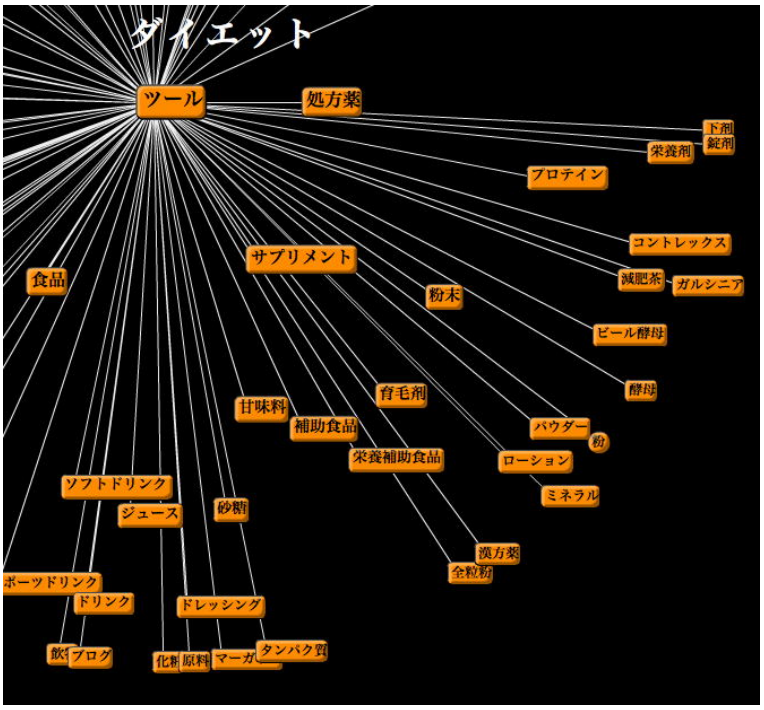

図 13: ダイエットの「ツール」カテゴリ

が，実は「砂糖ミルクダイエット」なるダイエット方法の一つに用いられているといった情報を得ることができます．

「準備」・「利用」ボタンを押すと，検索語についてその「準備」や「利用」を表す動詞（例えば，自動車「トゥーラン」の利用は「乗る」という動詞で表されます．）を表示させることができます．また，これらの表現の中には，トラブル，方法，ツールと関連の深いものがあり，そういった表現は黄色い枠で囲われています．そして黄色い枠で囲われた言葉をクリックすることで，その動詞と関連の深い，トラブル，方法，ツールが表示されます．図 14 は，上位概念「魚」を継承した検索語「アオブダイ」の利用表現「食べる」について，トラブル，方法，ツールを表示している例です．準備・利用表現については，上位概念を継承した場合に多く現れる傾向があります．

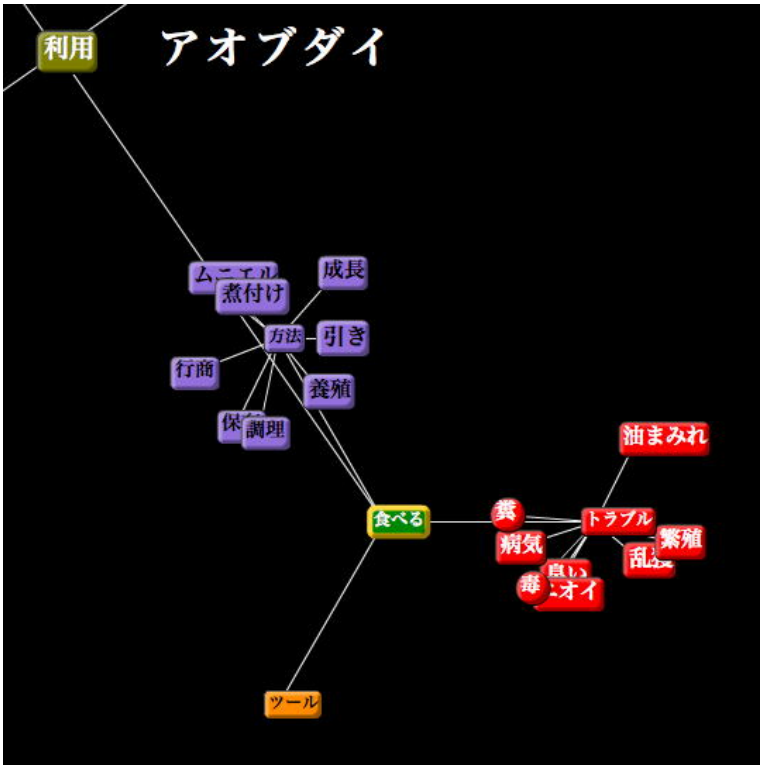

図 14: 準備・利用表現

謝辞

本研究を進めるにあたって，独立行政法人 情報通信研究機構 知識創成コミュニケーション研究センター 言語基盤グループ 鳥澤健太郎グループリーダー，村田真樹研究員には日頃から研究方針，研究内容など本研究に関して幅広くご指導頂いたことを，この場を借りて厚く御礼申し上げます．また，情報通信研究機構にて外部研究を行うにあたり，大学側から補助ならびに指導していただいた東条敏教授に深く感謝の意を表します．

参考文献

[1] 高橋哲朗, 岡本青史, 友澤大輔. ブログ上のクチコミ情報分析. 人工知能学会 知識流通ネットワーク研究会, 2008.

[2] 風間淳一, Stijn De Saeger, 鳥澤健太郎, 村田真樹. 係り受けの確率的クラスタリングを用いた大規模類似語リストの作成. 言語処理学会第 15 回年次大会, 2009.

[3] 野田陽平, 清田陽司, 中川裕志. 意外性のある知識発見のための wikipedia カテゴリ間の関係分析. 第 20 回セマンティックウェブとオントロジー研究会, 2009.

[4] 隅田飛鳥, 吉永直樹, 鳥澤健太郎, 萬成賢太郎. Wikipedia からの大規模な上位下位関係の獲得. 言語処理学会第 14 回年次大会, pp. 769–772, 2008.

[5] Stijn De Saeger, Kentaro Torisawa, Jun'ichiKazama. Looking for trouble. In *Proc. of The 22nd International Conference on Computational Linguistics (Coling2008)*, 2008.

[6] Keiji Shinzato, Tomohide Shibata, Daisuke Kawahara, Chikara Hashimoto, and Sadao Kurohashi. Tsubaki: An open search engine infrastructure for developing new information access. In *Proc. of IJCNLP*, pp. 189–196, 2008.